Evaluation of a Remote Data Collection Method to Study Human-Automation Interaction and Workload

Meghan Chandarana[1], Eric T. Chancey[2], Lisa R. Le Vie[2], and Michael S. Politowicz[2]

[1]NASA Ames Research Center, Moffett Field, California, USA
[2]NASA Langley Research Center, Hampton, Virginia, USA

Technological advances have increased the automation of Uncrewed Aerial Vehicles, allowing human operators to manage multiple vehicles at a high-level without the need to understand low-level system behaviors. Previous laboratory studies have explored the relationship between reliability, trust, use of automation, and the effects of number of vehicles under supervision on subjective workload. Due to limitations resulting from the COVID-19 pandemic, in-person laboratory studies are not always possible. Therefore, this work aimed to investigate if remote data collection alternatives, such as Amazon's Mechanical Turk, can provide comparative results as those obtained in laboratory settings. A study was conducted in the context of small drone operations. As expected, higher reliability led to higher trust ratings and the inclusion of more vehicles led to higher workload. In contrast, reliability unexpectedly had no significant effect on intention to use the automation. Though these results were encouraging, several limitations were identified.

INTRODUCTION

As Uncrewed Aerial Vehicles (UAVs) become increasingly more automated, human supervisory control roles are shifting from that of a single operator supervising a single vehicle to that of a single operator supervising multiple, highly automated vehicles. This one-to-many paradigm has been used across a variety of application domains including search and rescue (e.g., Chandarana et al., 2021; Liu, Ficocelli, & Nejat, 2015), foraging (e.g., Nam et al. 2018), and military operations (e.g., Chen & Barnes, 2012). An important aspect of enabling humans to effectively manage such highly automated vehicles is their trust in the automation. Past studies show that trust mediates the relationship between reliability and dependence and shows that future studies are needed to further understand the complex relationship (Chancey et al. 2015, 2017).

However, social distancing and precautions to prevent the spread of Coronavirus Disease 2019 (COVID-19) have disrupted research in a variety of domains (Servick et al., 2020). This disruption extends to human factors and applied experimental psychology research, which is generally conducted in laboratory settings, or *in situ,* using co-located human subject participants. Researchers, therefore, are beginning to pursue alternative avenues to collect meaningful data. For example, in response to COVID-19 restrictions, McLaughlin et al. (2020) present a review of methods for remote evaluation of medical devices to address usability and human factors issues (many of which are useful approaches beyond medical device evaluations specifically). Although McLaughlin et al. elaborated on multiple web-based services useful for remote data collection (e.g., Survey Monkey), they did not explore Amazon's Mechanical Turk (MTurk) service, a relatively new marketplace and tool for conducting remote data collection.

MTurk

MTurk is a crowdsourcing website, which has increasingly been used by the social sciences to post questionnaires, recruit and compensate participants, and store response data for subsequent analyses. Some research indicates MTurk data show comparable psychometric qualities as those conducted in the laboratory, such as reliability (Buhrmester et al., 2011) and external and internal validity (Berinsky et al., 2012). In relation to laboratory-based research, Rice et al. (2017) provides a practical guide to the advantages and disadvantages of using MTurk. Among the advantages Rice et al. notes are the relatively cheap and easy access to a diverse set of participants (i.e., not limited to college student populations), which may contribute to greater generalizability of results. Alternatively, some of the disadvantages indicated are poor response rates (i.e., starting but not completing a study), inattentiveness, participants rushing completion to gain compensation, and that online research is generally limited to measuring attitudes and perceptions (rather than behaviors).

As researchers explore remote data collection options, such as MTurk, it will become increasingly important to produce empirical evidence showing if experimental paradigms used in laboratory settings translate well to online environments. The purpose of the current study was to investigate the value of MTurk as an alternative approach to in-laboratory data collection. Specifically, we used MTurk to examine the effects of automation reliability on trust and intention to use automation in the context of monitoring highly automated drones. Additionally, the number of drones to monitor was also manipulated, to investigate its effects on subjective workload. If MTurk produces relatively similar results to those observed in laboratory settings (i.e., hypotheses are supported), then this would provide some evidence to indicate the value of MTurk as a remote data collection option for paradigms similar to the one outlined in this work.

Previous Research and Hypotheses

Functional reliability is defined by the number of errors an automated system produces during a given time period (e.g., a 90% reliable system produces 1 error for every 10 times it is engaged; Sullivan et al., 2008). Research has shown that higher functional reliability leads to higher automation use and, importantly, that there exists a "cross over point" of 70% reliability – below which the human may be better off without the automation (Wickens & Dixon, 2007). Intentions, based on attitudes (e.g., trust), lead to behaviors (Ajzen & Fishbein, 1977, 1980). Studies have shown that intention to use technology (e.g., automation) is predictive of actual technology use (e.g., Davis et al., 1989; Venkatesh et al., 2003). Similarly, therefore, higher reliability should lead to higher intention to use automation. Moreover, higher reliability has been generally shown to lead to higher trust in automation (see Lee & See, 2004, for theoretical perspective; for empirical evidence see Chancey et al., 2015, Chancey et al., 2017). These well-established relationships led us to hypothesize the following:

- Hypothesis 1: higher automation reliability will lead to higher intention to use the automation
- Hypothesis 2: higher automation reliability will lead to higher trust in the automation.

Research has also shown that subjective workload ratings are sensitive to the number of vehicles being supervised, where increasing the number of vehicles leads to higher workload ratings (Cummings et al., 2014; Galster et al., 2001). Additionally, observers can track a maximum of about 4 moving objects (see Alvarez & Franconeri, 2007 and Tran & Hoffman, 2016). These findings from the literature led us to our third hypothesis.

- Hypothesis 3: More drones will lead to higher subjective workload ratings.

METHOD

Participants

The experiment included *n* = 90 (35 females) participants, with a mean self-reported age of 40 years (*SD* = 11.96). Eligible participants were 18 years or older, had a "Master Qualification" (defined below) status in MTurk, and currently resided in the United States (US). Participants received $1.81, which was based on the US federally-set minimum wage ($7.25) and an estimated 15 minutes to complete the study (average completion time was 12 m, 34 s). Participants completed the study only once.

MTurk Task Overview

With MTurk, "Requestors" (e.g., researchers) can recruit, hire, and compensate "Workers" (e.g., participants) to complete a variety of tasks (e.g., complete surveys). Each study is labeled a "Human Intelligence Task" (HIT), which represents individual jobs that Workers can respond to and are compensated for completing. Once completed, HITs are either approved or rejected for satisfactory completion by Requestors (i.e., participants are either approved and compensated, or rejected and not compensated). Workers are associated with specific "System Qualifications" assigned to them by MTurk as well as "Customized Qualifications" that are assigned by Requestors. Requestors can use these qualifications to filter out workers that are eligible or ineligible to participate in their study. One example of a qualification filter is the "Master Qualification," which is a title assigned by MTurk that represents Workers who have demonstrated a high level of consistently and adequately completing assignments (e.g., a study) across a range of Requestors. Examples of qualifications that may be selected as filters are percentage and number of approved HITS, geographic location, gender, level of education, and work experiences, among others.

Design

The experiment employed a 3 × 3 between-subjects design. The study design manipulated vehicle reliability (60%, 90%, or 100%) and number of vehicles to monitor (1, 4, or 8). For this study, stated reliability measures were used only, like the study presented by Bliss et al. (1995).

Dependent Measures. *Human-automation trust* was measured with a modified version of Madsen and Gregor's (2000) questionnaire used in Chancey et al. (2017). Due to the remote and non-interactive nature of the study, a behavioral measure for automation use could not be used. Therefore, a modified version of Rahman et al.'s (2017) questionnaire was used to measure *Intention to Use Automation*. These measures showed adequate internal consistency with $\alpha_{Cronbach's} > .70$. For these measures, a statement was presented and participants indicated their agreement from Strongly Disagree (1) to Strongly Agree (100) on a slider. *Subjective Workload*

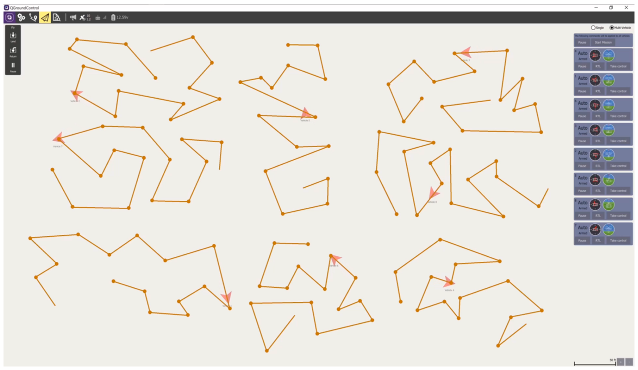

Figure 1: UAV monitoring interface used during remote data collection. This example shows 8 vehicles and their respective waypoints and flight paths.

was measured via the NASA Task Load Index (TLX; Hart & Staveland, 1988).

Procedure

For this study, participants watched an experimentally manipulated video depicting highly automated small drone operations (Figure 1) and then responded to a series of questionnaires. Participants self-selected into the study (hosted on SurveyMonkey.com) via MTurk and were compensated $1.81 for taking part in the study. If participants agreed to take part in the study, they were randomly assigned to 1 of 9 experimentally manipulated videos depicting highly automated small drone operations (see videos that correspond to group codes in Figure 2).

Group Code	1 Drone	4 Drones	8 Drones
60% Reliable	1	2	3
90% Reliable	4	5	6
100% Reliable	7	8	9

Figure 2: Group codes to match videos

Videos:

- Group 1: https://tinyurl.com/r9fx5z62
- Group 2: https://tinyurl.com/2p89h6yn
- Group 3: https://tinyurl.com/2p977ctv
- Group 4: https://tinyurl.com/27cr7m82
- Group 5: https://tinyurl.com/y2r68u3x
- Group 6: https://tinyurl.com/4h44jz43
- Group 7: https://tinyurl.com/2p8ejjh5
- Group 8: https://tinyurl.com/yckkafzn
- Group 9: https://tinyurl.com/ysas3zf2

Following the video, participants filled out a series of questionnaires (items randomized): Post-Video Items, Human-Automation Trust, Perceived Risk, Subjective Workload, Intention to Use Automation, Automation Complacency Potential, Propensity to Trust Automation, Self-Efficacy Scale, and Demographics (note: only the Human-Automation Trust, Subjective Workload, and Demographics questionnaire results are presented in this work).

Several approaches were taken to promote data quality (see Cheung et al., 2017). Data quality "item-checks" were placed throughout the questionnaires (e.g., "If you are paying attention, select a number in the twenties"). Participants that incorrectly answered any item-check question were excluded from analyses. An outlier-labeling rule with a multiplier of 2.2 was consulted to identify lower-limit completion times (Hoaglin & Iglewicz, 1987). Participants that completed the entire experimental session quicker than 6 m, 57 s were excluded from the analyses. A benign warning was included on the participant agreement screen: "I understand that my Human Intelligence Task (HIT) may also be rejected and payment forfeit if I fail to follow instructions or provide adequate responses according to the directions. Specifically, response patterns will be monitored and any indications of random responding will result in no compensation." The study was purposefully limited in time and scope to curb participants abruptly quitting or responding randomly (Rice et al., 2017). Finally, a progress indicator bar at the bottom of the screen was used to reduce the rate of participants quitting abruptly due to a lack of awareness for progression and screens remaining to complete (Rice et al., 2017).

RESULTS

Data were inspected for outliers and equal numbers among groups. To minimize the chances of making a Type I error, $p < .05$ was established to indicate statistical significance.

Intention to Use Automation

Failing to support hypothesis 1, a significant main effect of reliability on intention to use automation was not observed ($p > .05$). No significant main effect of the number of drones on intention to use the automation was observed either.

Human-Automation Trust

Supporting hypothesis 2, there was a significant main effect of reliability on trust in the automation, F (2, 81) = 4.63, p = .012, *partial* η^2 = .10, observed power = .76. Participants in the 60% reliable group (M = 70.30, SD =

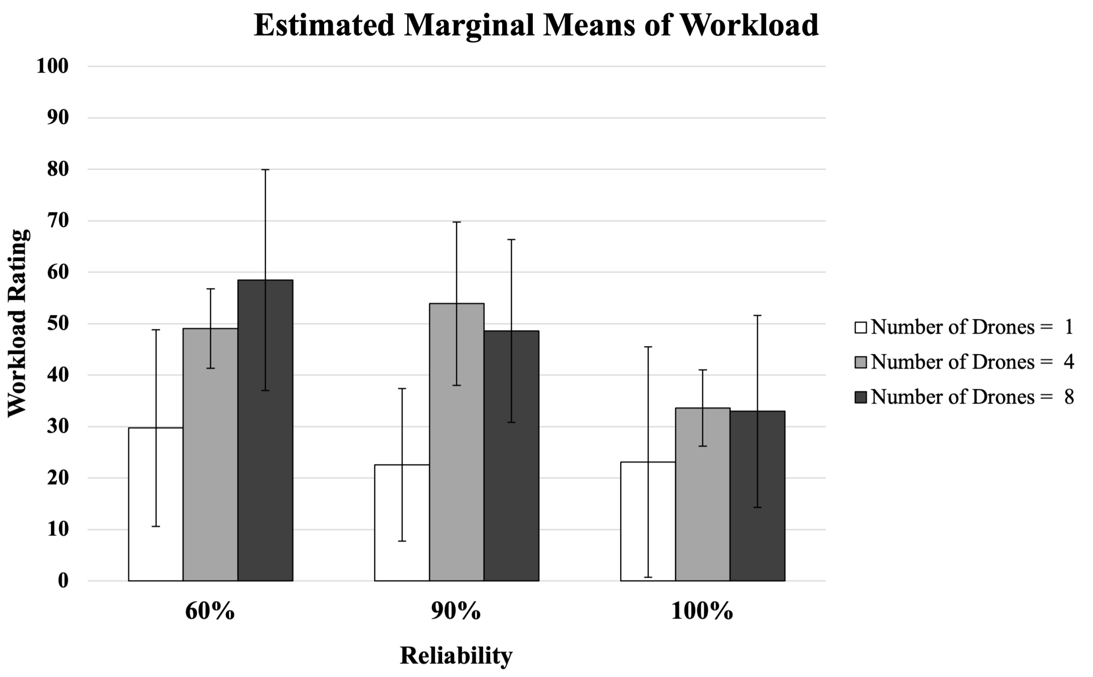

Figure 3: Effects of reliability and number of drones on workload.
Note: Error bars represent standard deviations.

26.78) trusted the Autopilot significantly less than the 90% reliable group (p = .046, M = 84.07, SD = 15.40) and 100% reliable group (p = .021, M = 85.73, SD = 22.01). The main effect of number of drones on trust was not statistically significant, F (2, 81) = 2.67, p = .075, *partial* η^2 = .06, observed power = .56.

Subjective Workload

Supporting hypothesis 3, there was a significant main effect of number of drones on subjective workload, F (2, 81) = 13.82, p < .001, *partial* η^2 = .25, observed power = .99 (Figure 3). Participants in the 1 drone group rated their workload (M = 25.14; SD = 17.80) significantly less than both the 4 drone group (p < .001; M = 45.52; SD = 18.68) and 8 drone group (p < .001; M = 46.66; SD = 20.64). Additionally, there was a significant main effect of reliability on subjective workload, F (2, 81) = 6.39, p = .003, *partial* η^2 = .14, observed power = .89. Participants in the 100% reliability group (M = 29.89; SD = 20.28) rated their workload as significantly less than both the 90% reliability group (p < .001; M = 41.67; SD = 17.89) and 60% reliability group (p < .001; M = 45.76; SD = 22.88).

DISCUSSION

Encouragingly, higher reliability led to higher trust ratings (supporting Hypothesis 2) and increasing the number of drones led to higher subjective workload ratings (supporting Hypothesis 3). However, the effect sizes for the results supporting Hypothesis 2 were notably smaller than those observed in some laboratory-based studies (e.g., Chancey et al. 2015, 2017). A significant effect of reliability on intention to use the automation was not observed (failing to support Hypothesis 1). The lack of support for Hypothesis 1 was likely due to the method employed (i.e., substituting an intention for a behavior), as there is a significant body of evidence supporting the relationship between reliability and behavioral automation use in lab-based studies (see Wickens & Dixon, 2007).

In addition to the smaller effect sizes, several limitations resulted from the remote data collection environment. First, participants were limited to only an observation/monitoring type task. MTurk studies are conducted asynchronously without any direct interaction with study material, therefore tasks that require real-time interaction with the task interface were not possible (cf. Rice et al., 2017). This limitation inherently reduces the scope of metrics that can used in such remote studies (e.g., needing to measure intention to use the automation rather than measuring automation use directly).

Another limitation of the remote data collection environment is the inability to control external data collection reliability. Although several methods were employed to ensure data quality (as detailed above), researchers have no direct control over the surroundings of the participants when they are completing the given tasks. There is no way to know if outside influences or distractions are reflected in the data collected or if those data would reflect the outcome of similar situations in the real-world environments these vehicles are expected to be fielded in.

Though some general similarities between MTurk and laboratory-based studies can be derived from this work, a more definitive method would be to replicate this study in a laboratory. Yet, to the premise of this study (see Introduction), at the time of data collection lab-based research was not possible due to COVID-19 restrictions.

CONCLUSION

Overall, using Amazon's MTurk as a remote data collection environment produced encouraging results. In general, the data showed similar trends previously seen in laboratory environment studies. However, the effect sizes from the MTurk study were much smaller. In addition, several limitations exist inherently with remote data collection: (1) inability to allow participants to complete tasks that require direct interaction with the interface and (2) no direct control over the environment of the participants when completing tasks. Although these results may be encouraging for additional remote-data collection activities, additional work needs to be completed to increase confidence in this approach.

This work was supported by NASA's Autonomous Systems Enduring Discipline under the Transformational Tools and Technologies Revolutionary Aviation Mobility Subproject. The views expressed are those of the authors and do not necessarily reflect the official policy or position of NASA or the U.S. Government.

REFERENCES

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin*, *84*(5), 888-918.

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Upper Saddle River, NJ: Prentice Hall.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7*(13):14, 1–10.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Political Analysis, 20, 351-368.

Bliss, J. P., Dunn, M., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. Perceptual and Motor Skills, 80, 1231-1242.

Buhrmester, M., Kwang, T., & Gosling, S. D., 2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6 (1), 3-5.

Chancey, E.T., Bliss, J. P., Proaps, A. B., & Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors, 57* (6), 947-958.

Chancey, E.T., Bliss, J.P., Yamani, Y., & Handley, H.A.H. (2017). Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors, 59* (3), 333-345.

Chandarana, M., Hughes, D., Lewis, M., Sycara, K., & Scherer, S. (2021). Planning and Monitoring Multi-Job Type Swarm Search and Service Missions. *Journal of Intelligent & Robotic Systems, 101*(3), 1-14.

Chen, J. Y., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. Human Factors, 54(2), 157-174.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology. An evaluation and practical recommendations. Journal of Business Psychology, 32, 347-361.

Cummings, M. L., Bertucelli, L. F., Macbeth, J., & Surana, A. (2014). Task versus vehicle-based control paradigms in multiple unmanned vehicle supervision by a single operator. *IEE Transactions on Human-Machine Systems, 44* (3), 353-361.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35* (8). 982-1003.

Galster, S. M., Duley, J. A., Masalonis, A. J., & Parasuraman, R. (2001). Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. *The International Journal of Aviation Psychology, 11*(1), 71-93.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hoaglin, D.C., & Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labeling. *Journal of American Statistical Association, 82*, 1147-1149.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46* (1), 50-80.

Liu, Y., Ficocelli, M., & Nejat, G. (2015, October). A supervisory control method for multi-robot task allocation in urban search and rescue. In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)* (pp. 1-6). IEEE.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. Proceedings of the Eleventh Australian Conference on Information Systems.

McLaughlin, A. C., Drews, F. A., Vaugn-Cooke, M., Kumar, A., Nesbitt, R. R., & Cluff, K. (2020). Evaluating medical devices remotely: Current methods and potential innovations. Human Factors, 62 (7), 101-1060.

Nam, C., Li, H., Li, S., Lewis, M., & Sycara, K. (2018). Trust of humans in supervisory control of swarm robots with varied levels of autonomy. In *2018 IEEE international conference on systems, man, and cybernetics (smc)* (pp. 825-830). IEEE.

Rahman, M. M., Lesch, M. F., Horrey, W. J., & Strawderman, L. (2017). Accident Analysis and Prevention, 108, 361-373.

Rice, S., Winter, S. R., Doherty, S., & Milner, M. (2017). Advantages and disadvantages of using internet-based survey methods in aviation-related research. Journal of Aviation Technology and Engineering, 7 (1), 58-65.

Servick, K., Cho, A., Gugliemi, G., Vogel, G., & Couzin-Frankel, J. (2020). Updated: Labs go quiet as researchers brace for long-term coronavirus disruptions. *Science Magazine*. Retrieved 2 February, 2022 from https://www.science.org/content/article/updated-labs-go-quiet-researchers-brace-long-term-coronavirus-disruptions.

Tran, A., & Hoffman, J. E. (2016). Visual attention is required for multiple object tracking. *Journal of Experimental Psychology: Human Perception and Performance*.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8* (3), 201-212.

Venkatesh, V., Moris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27* (3), 425-478.